Integrating a Natural Language Message Pre-Processor with UIMA

Eric Nyberg, Eric Riebling,
Richard C. Wang & Robert Frederking

Language Technologies Institute
Carnegie Mellon University

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Integrating a Natural Language Message Pre-Processor with UIMA	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Language Technologies Institute,5000 Forbes Avenue,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Workshop ?Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP? at the Sixth Conference on Language Resources and Evaluation (LREC-2008). Marrakech, Morocco. 2008

14. ABSTRACT

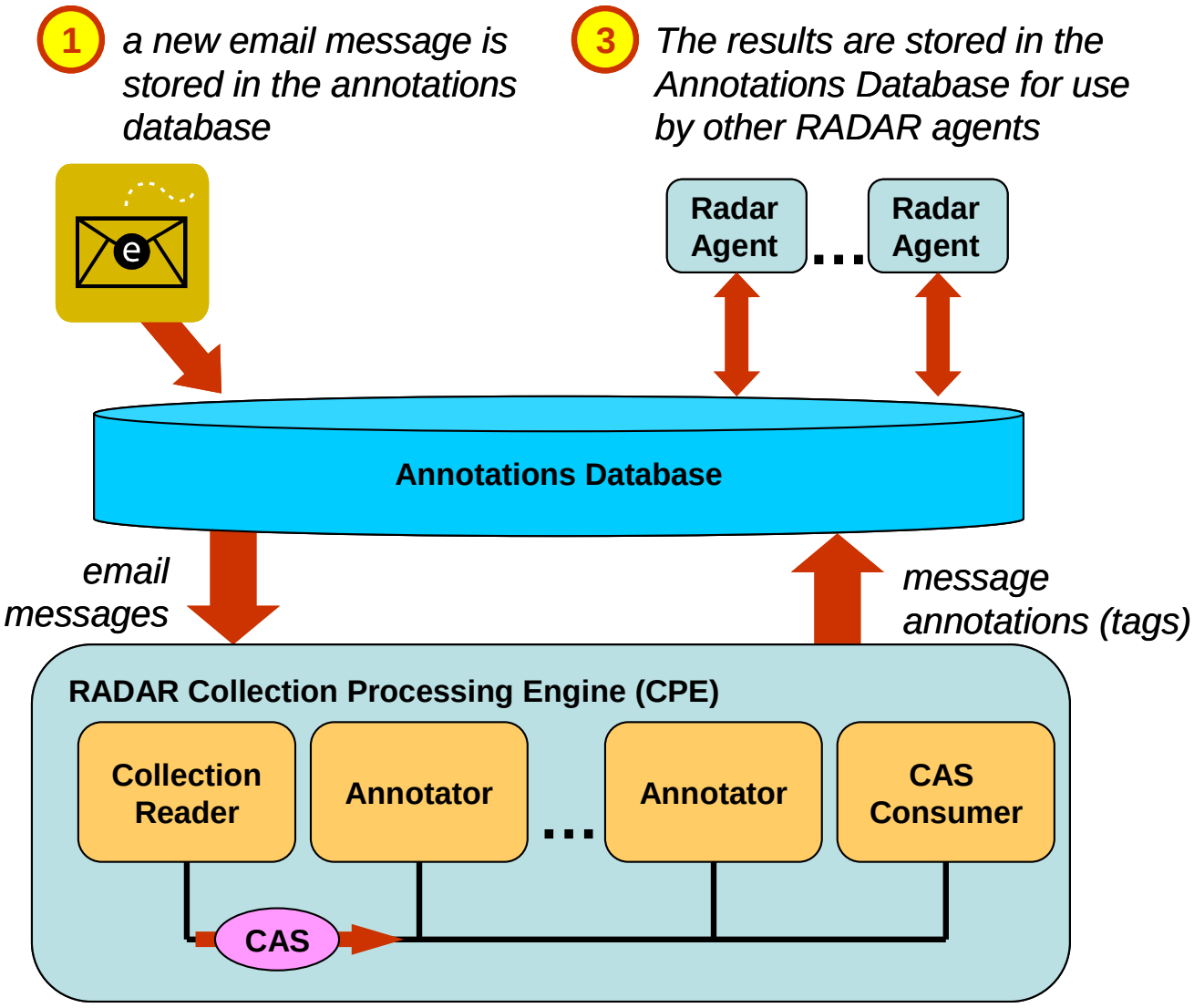
15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**16**	

Project Context

- RADAR: Goal is to help desktop user
  - Personal Assistant that Learns (PAL)
  - Test environment: conference planning
  - Primary input: email messages!
- Requirements include:
  - Preprocessing of email messages
    - Segmentation, typo correction, etc.
    - Syntactic parsing
    - General and domain-specific semantic interpretation
    - Domain-specific task request extraction
  - Original content preserved

The RADAR CPE



Radar Annotator implementation

- Three varieties of implementation:
  - MinorThird toolkit (William Cohen)
  - Java code
  - Client/server (primarily for legacy and external vendor software) with UIMA wrapper for client

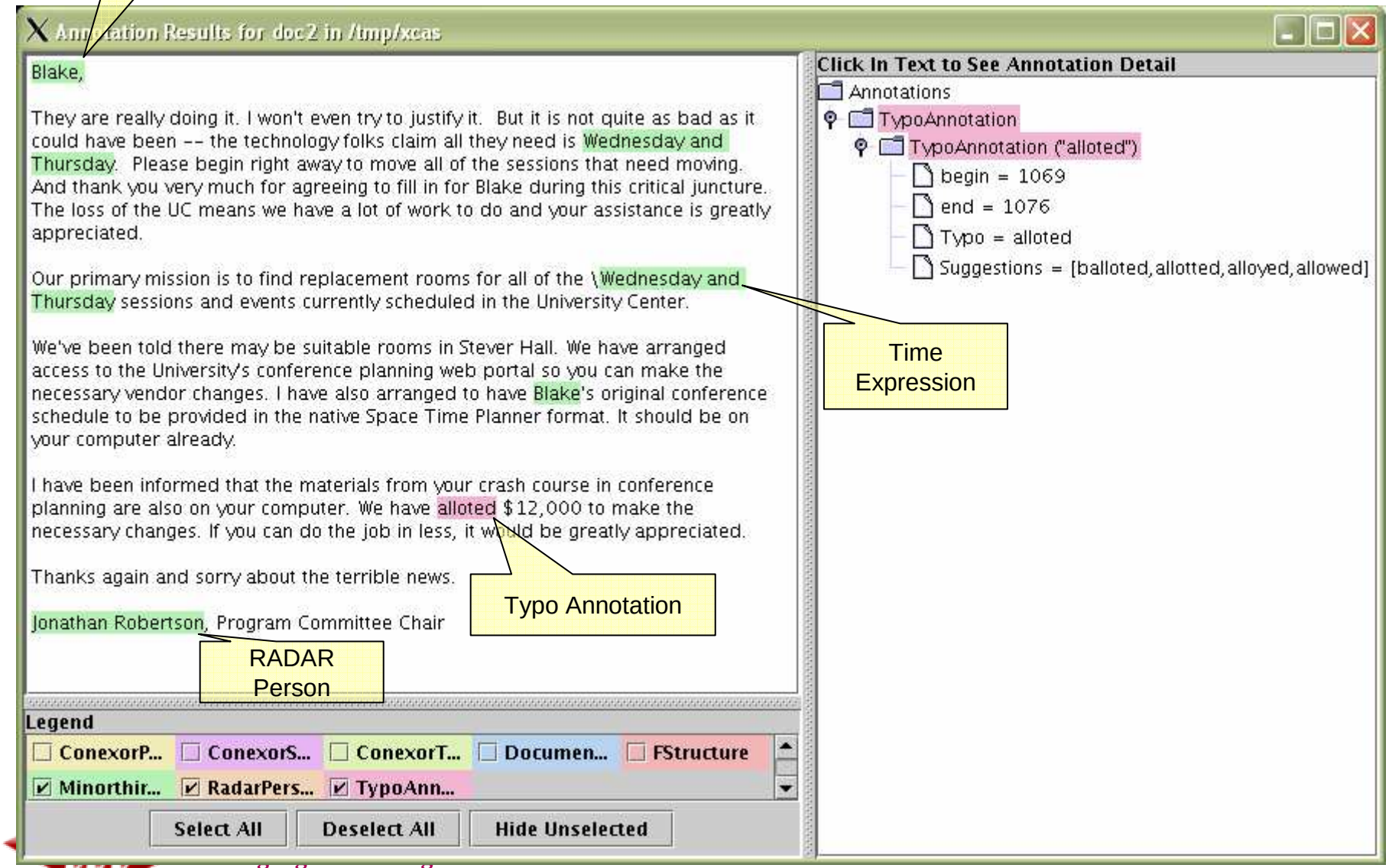
Radar Annotators: list

- In order:
- Collection Reader
  Email Opening
  Conexor parser
  Temporal Expr.
  F-Structure
  GFrame
  DFrame

- In between:
- Task
  RADAR Person
  SCONE Semantics
  Person Name
  SCONE Implicit
  Space Request
  Typo

- Last:
- CAS Consumer

Example Document



Sample Annotations: TempEx

String	Offset	Length
of the summer	73	13
This summer	175	11
three days	359	10
1 week	493	6
Starting May 10)	774	16
July 4	1971	6
INDEPENDENCE DAY	1939	16

Sample Annotations: Typo

String	Offset	Length	Value
teh	60	3	the, eh,…
brousing	20	8	rousing, browsing
bris	16	4	…brisk...
midle	36	5	middle,…
infor	117	5	inform,…
committe	83	8	committed
fed	286	3	fled

Sample Annotations: DFrame

String	Offset	Len	Value
Which room is the first event in?	14	33	((dframe … (subj ((POS N) (attr ((POS NUM) (function attr) (ortho first) (root first) …

Sample Annotations: BriefingReq

String	Offset	Len	Value
I need a progress report on yesterday NOW	0	43	<node id="request11 72260778347" … </node>
please send me a campus map sooon. -- chian	0	44	<node id="request11 72261238858" … </node>

%	Time(ms)	s/doc	Annotator
65.27	5310311	21.24	DFrame
24.60	2001145	8.00	GFrame
			RADAR Person
			SCONE Sem.
			Temporal Expr.
1.03	83563	0.33	Person Name
0.71	57742	0.23	SCONE Impl.
0.54	44187	0.18	F-Structure
0.18	14889	0.06	Email Opening
0.17	13513	0.05	SpaceRequest
0.17	13445	0.05	Conexor
0.07	5835	0.02	Typo
0.06	4746	0.02	CAS Consumer
0.03	2725	0.01	Collection Reader
0.03	2415	0.01	Task
100.00	8136349	32.55	Entire Pipeline

Expensive rule-based computations (structural transformation rules and KB lookups)

[sample: 250 randomly selected messages]

7. Add domain semantics
6. Add general semantics
Label known person names
Domain KB interpretation
4. Add anchored time labels
Label any person name
Add domain KB features
5. Label grammatical roles
2. Label salutations in email
Label space requests
3. Segmentation, parsing
Label typo fixes
8. Write to ADB
1. Read from ADB
Label task requests

Annotator Run-Time

Carnegie Mellon
Language Technologies Institute

Sample Annotator Precision

Annotator	% Correct	% Partly Correct
Vendor Order Annotator	100%	--
Task Annotator	73%	77%
Person Name Annotator	76%	85%
Space Request Annotator	64%	79%

[sample: 50 randomly selected messages]

Since the RADAR context is machine assistance in a human task, these should also be correlated with their effect on human task performance (currently assessed end-to-end for full system).

Cost of Adoption

- 1.5 months FTE to wrap and integrate 15 NLP components
  (programmer already familiar with UIMA)

Issues/Future Work

- Better robustness/decoupling
  - Require standard service interfaces for NLP components
  - Wrap as UIMA-EE (UIMA-AS) services
- Better transparency
  - Hard to tell whether a service is dead or just working hard
  - Need better logging/communication with services
- Better speed
  - Optimize rule-based engines
  - Provide multiple service instances for time-consuming services

Questions?

References

Cohen, William W. (2004). Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net.

Han, Benjamin, Donna Gates and Lori Levin (2006).Understanding temporal expressions in emails. *Proceedings of the Human Language Technology Conference*, Association for Computational Linguistics.

Kumar, M. et al. (2007). Summarizing Non-textual Events with a 'Briefing' Focus. *Proceedings of RIAO*, Centre De Hautes Etudes Internationales D'Informatique Documentaire.

Nyberg, E., T. Mitamura, K. Baker, D. Svoboda, B. Peterson and J. Williams (2002). "Deriving Semantic Knowledge from Descriptive Texts using an MT System", *Proceedings of AMTA 2002*.

Yang, Y. et al. (2005). Robustness of Adaptive Filtering Methods in a Cross-Benchmark Evaluation. *Proceedings of ACM SIGIR*, 98–105. ACM Press.